Load Check Process

There are 3 basic checks:

1. Checking for all the records in Data Lake.

Source type	Formula	Filter criteria
callcenter, dns, donorlist, experian, marketinglist	Number of records in archived files = number of records in cleansed parquet files + number of records in error csv files (if any).	S3 Naming convention based on dataSource, Customer, BatchName and outputType(cleansed/error/archive)
utilization	Number of records after joining all the files in the archive batch folder = number of records in cleansed parquet files + number of records in error csv files (if any).	S3 Naming convention based on dataSource, Customer, BatchName and outputType(cleansed/error/archive)

2. Checking for all the records in Data Reservoir.

Source Type	Formula	
All sources	Number of records in cleansed parquet files = Number of records in datahub.cleansed_activities.	DataLake: BatchName = DataReservoir(cleansed_activities) : BatchId DataLake: Customer = DataReservoir(cleansed_activities): Customer

3. Checking for all the records in Activity Pipeline.

Source Type	Formula	Filter criteria
All sources	Number of records in cleansed parquet files = Number of records in datahub.person_activity.	cleansed_activities: BatchName = person_activity: BatchId cleansed_activities: Customer = person_activity: Customer